# Cosmic Ray Detection in Astronomical Images via Dictionary Learning and Sparse Representation

Srinadh Reddy Bhavanam, Sumohana S. Channappayya, Srijith P.K and Shantanu Desai
*Indian Institute of Technology, Hyderabad, Kandi, Telangana-502285, India*

***Abstract*—In this work, we propose a novel Dictionary Learning (DL) based framework to detect Cosmic Ray (CR) hits that contaminate the astronomical images obtained through optical photometric surveys. The unique and distinguishable spatial signatures of CR hits compared to other actual astrophysical sources in the image motivated us to characterize the CR patches uniquely via their sparse representations obtained from a learned dictionary. Specifically, the dictionary is trained on images acquired from the Dark Energy Camera (DECam) observations. Next, the learned dictionary is used to represent the CR and Non-CR patches (e.g., each patch is with $11 \times 11$ pixel resolution) extracted from the original images. A Machine Learning (ML) classifier is then trained to classify the CR and Non-CR patches. Empirically, we demonstrate that the proposed DL-based method can detect the CR hits at patch level and provide approximately $83\%$ detection rates at 0.1 $\%$ false positives on the DECam test data with Random Forest (RF) algorithm. Further, we used the coarse segmentation maps obtained from the classifier output to guide the deep-learning-based CR segmentation models. The coarse maps are fed through a separate channel along with the contaminated image to detect the CR-induced pixels more accurately. We evaluated the performance of proposed DL-guided deep segmentation models over the baseline on test data from DECam. We demonstrate that the proposed method provides additional guidance to the baseline models in terms of faster convergence rate and improves CR detection performance by $2\%$ in the case of shallow models. We made our dataset and models available at https://github.com/lfovia/Dictionary-Learning-Augmented-Cosmic-Ray-Detection.**

***Index Terms*—Cosmic ray hits, observational astronomy, image processing, dictionary learning, approximate K-SVD, sparse coding**

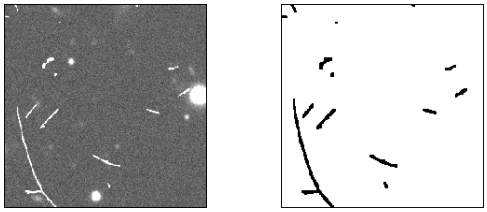
## I. Introduction

Wide-field optical and spectroscopic imaging surveys have yielded a wealth of astronomical data, allowing for a better understanding of the processes that drive the formation and evolution of the Universe and its contents. Future surveys, such as the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST; [1]), will contribute to our understanding of the Universe by extending observations to faint astronomical systems. Extracting accurate source catalogues from images of these surveys is crucial for a wide range of astronomical research areas. However, the source detection algorithms are limited in crowded fields or when images are contaminated by the detector, optical, or environmental imperfections. Imaging data obtained through astronomical surveys are often contaminated, resulting in spurious detection. By correctly identifying and masking the contaminants, it is possible to reduce the frequency of spurious detections in astronomical catalogues. Moreover, the sheer volume of data produced by modern wide-field surveys makes a visual inspection of contaminants impossible in most cases. Hence, developing fully automated methods to detect the contaminants over actual astrophysical sources is required in the image processing pipeline.

The CR hits illustrated from Fig. 1 are the most dominating and frequently occurring contaminants in astronomical observations involving solid-state detectors such as Charge-Coupled Devices (CCDs). These are related to high-energy particles which travel through space and the atmosphere until they hit the telescope optics. They affect the detector by accumulating excess charge in the affected regions compared to the source or background pixels [2]. CR hits in the optical images appear as bright and sharp patterns as they are not blurred by telescope or the atmosphere, similar to astrophysical sources. In addition, the CR hits appear in patterns like dots, lines or curves depending on their incidence angle with the detector. Moreover, a few CR events that look similar to faint astronomical sources sometimes is even more challenging. As a result, before further analysis of astronomical data, rejection of CR hits should be performed to guarantee that only high-quality data is used for scientific studies.

It is well known in the literature that the CR hits are transient in nature. This implies that the probability of the same pixel being contaminated (by CR hits) over multiple exposures of the same region of interest is relatively low. Hence, obtaining multiple exposures is one possible solution to the CR detection problem. With the availability of multiple successive exposures of the same sky region, the CR hits can be easily detected by discarding outliers within the same sky region [3]. However, obtaining multiple exposures is not always possible as it depends on the observation strategy adopted. Also, if one is looking for moving astrophysical sources, then the methods used for CR rejection using multiple exposures cannot be applied. Therefore, detecting and masking CR hits from single-exposure optical images is paramount.

In this work, we devised a novel Dictionary Learning (DL) based solution to detect the CR contaminated pixels in single exposure CCD data. The optical images typically constitute actual astrophysical sources like stars, quasars, galaxies, the contamination (including CR hits, hot/bad pixels, satellite trails, etc.), and the background. Contrary to the CR hits mapped in patterns like dots, lines, and curves, the sources are presented as point sources (stars and quasars) or extended sources (galaxies). Hence, the CR hits are easily distinguishable from the actual astrophysical sources through their unique spatial signatures. Dictionary learning [4] which learns a set of



(a) Image with CR hits (b) Groundtruth CR mask

Fig. 1: CR contaminated image with 256 $\times$ 256 cutout.

image features for robust image representation, is widely used in several image processing tasks, including image denoising [5], recognition [6], [7], etc. The unique spatial signatures of CR hits and the feature representation capability of the DL method motivate us to represent the CR patches via their sparse representations obtained from a learned dictionary and thus separate them from actual astronomical sources. Our primary contributions mainly include:

- We propose a novel DL-based CR detection model on the DECam imager. We first create patches from the original images and are characterized via sparse representation obtained from a learned dictionary. The dictionary is learned from the patches extracted from the original images. Classical ML algorithms are then used to distinguish between CR and Non-CR patches. The classifier output provides a coarse CR segmentation map using patches.
- In addition, coarse CR maps obtained from the proposed DL algorithm are fed to the Convolutional Neural Network (CNN) based models (such as deepCR [8]) through a separate channel to provide additional supervision to the deep-learning-based CR segmentation models.

The rest of this paper is organized as follows. In Section II, related work on CR detection from single exposure CCD data is presented. The details on data collection and synthesis are presented in Section III. Section IV and Section V presents the proposed methodology and summary of the results, respectively. Finally, conclusions are marked in Section VI.

## II. Related Work

The literature presents numerous methods to detect the CR hits in single exposure CCD and spectroscopic images [9]–[12]. Among these methods, LACosmic [12] and it's multi-core optimized implementation, Astro-SCRAPPY [13], showed state-of-the-art performance. LACosmic employs an iterative approach to detect the CR induced pixels in an image using a laplacian-like edge detection kernel. Furthermore, because LACosmic is an iterative technique, CR detection takes a long time. Supervised Machine-Learning (ML) classification algorithms such as K-nearest neighbours, multi-layer perceptrons [14], and decision-tree algorithms [15] were also used for CR detection and yielded encouraging results on small Hubble Space Telescope (HST) datasets. However, the generalization performance of these algorithms is relatively low when compared to LACosmic and Astro-SCRAPPY.

Compared to the hand-crafted kernels, kernels learned using CNNs through backpropagation show better performance in several image processing tasks. Masking of CR hits is also performed with deep CNNs in recent years [8], [16], [17]. deepCR [8] (U-Net [18] based architecture) is the initial attempt towards using deep learning and provide best in class solution for CR detection on spatial HST ACS/WFC F606W images. On the other hand, MaxiMask [16] is provided for ground-based imaging surveys to detect several contaminants, including CR hits. MaxiMask employed images from multiple instruments and ground-based observation sites to ensure that the model covers the most recent astronomical surveys. By addressing issues like class imbalance (between CR and Non-CR pixels) and developing individual models for different instruments, Cosmic-CoNN [17] attempted to produce generic models and demonstrated decent results.

## III. Dataset

Images from the Dark Energy Camera (DECam) [19] instrument are used in this work. The DECam is a 570-megapixel camera with a 2.2-degree field of view and a pixel scale of 0.263 arcsecond/pixel. It is housed at the Cerro Tololo Inter-American Observatory (CTIO) on the Victor M. Blanco 4-meter Telescope. For training and testing the CR detection models, raw scientific images from four different photometric bands of DECam are used, namely $g$, $r$, $i$, and $z$, each having a 90 sec exposure period. These images, which have a resolution of $4K \times 2K$ pixels, are made up of DECam detrended data from the Science Verification phase, which spanned from November 2011 to February 2012, and were processed using the `CosmoDM` pipeline [20], [21], which removes CR hits using the algorithm outlined in [3]. Similar to [16], we have also used dark data (obtained when no light is fallen on the detector, and so the only contributors to the content of undamaged pixels are the offset, dark current, noise and CR hits) from DECam to generate CR contamination synthetically.

## IV. Proposed Method

The CR hits look spatially different from other actual astrophysical sources in the image, and even they appear as sharp and bright patterns. Several classical algorithms leverage using one or more of these features to detect the CR hits in astronomical images. Similarly, in this work, we consider the spatial signatures of CR hits, which show that they appear in patterns like dots, lines, and curves illustrated from Fig. 1 as their unique features. Then we propose to characterize the CR patches via their sparse representations obtained from DL. Sparse and redundant representation assumes that natural signals can be described as a linear combination of a few atoms from the pre-defined dictionary. Specifically, the dictionary is learned on image patches obtained from the DECam observations. In astronomical images, Non-CR patches (including sources and background) are the natural and desired signals, and CR hits can be considered as anomalies.

Once the dictionary is learned, it is used to represent both the CR and Non-CR patches. Then, some classifier is trained to distinguish between CR and Non-CR patches. The classifier helps to provide coarse CR segmentation maps from its output as we use patches here rather than pixels. Further, we used the coarse CR segmentation maps obtained from the proposed DL approach to provide additional supervision to the CNN-based CR segmentation models like deepCR [8] to improve the CR detection performance. Complete details on these approaches are described as follows:

### A. Patch Classification (Coarse CR Segmentation)

*1) Dictionary Learning and Sparse Representation:* Given training data *X* with *N* vectors each of dimension *P* denoted by $x_n \in \mathbb{R}^P$, the dictionary learning aims at obtaining a basis set called dictionary $D \in \mathbb{R}^{P \times K}$. The dictionary is used to approximate the data via a linear combination of a few basis vectors from its columns (atoms). The sparsity is obtained by multiplying the dictionary matrix *D* with another sparse code matrix $Y \in \mathbb{R}^{K \times N}$. This can be accomplished by the constrained optimization as follows [4], [22], [23]:

$$\min_{D,Y} ||X - DY||_F^2 \quad \text{Subject to } \forall i \; ||y_i||_0 = M \tag{1}$$

The problem from Eq. 1 can be solved effectively using the Approximate K-SVD algorithm [23]. Approximate K-SVD provides an over-complete dictionary that is learned from training data *X* with *K* atoms. Once the dictionary is learned, the new samples are represented using this as a *K* dimensional vector. For this, we used Orthogonal Matching Pursuit (OMP) [24] algorithm to make this representation sparse by restricting *M* to be the number of non-zero coefficients.

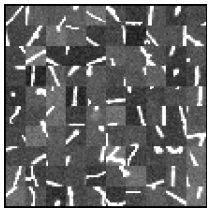
*2) Machine Learning Algorithms:* Once the representation is obtained via learned dictionary, we used the classical ML classifiers such as Random Forest (RF) [25] to discriminate the patches between CR and Non-CR. The classifier provides coarse CR map from it's output.

### B. Dictionary Learning Guided CNN Model

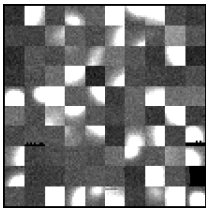
The coarse CR maps obtained through the proposed DL-based model are used to guide the CNN-based deep segmentation models. We considered the deepCR [8] model as the baseline in this work. The input to the proposed DL-guided CNN-based model comprises the original image plus another channel containing coarse CR maps. While for the baseline, only the original contaminated image is fed. The baseline and proposed DL-guided models are U-Net-based architectures with encoder-decoder paths and skip connections.

## V. Results and Discussion

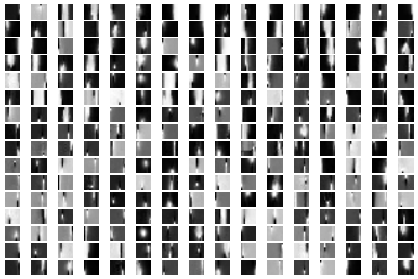
For training and evaluation of the proposed methods, we used 56 raw science images from each photometric band *'griz'* of the DECam instrument. These images are originally with the resolution of 4K $\times$ 2K. Out of these 56 images per band, 40 are used for training, and 16 are used for evaluation. Below is a detailed description of how we used this dataset and the performance of the proposed CR detection models.



(a) CR Patches



(b) Non-CR Patches



(c) Dictionary atoms learned on Non-CR Patches

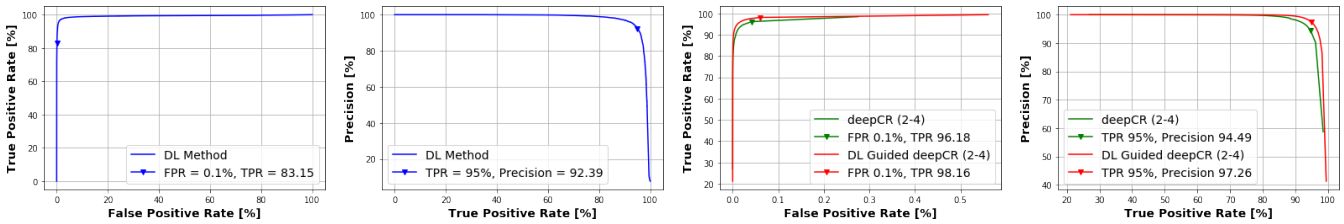
Fig. 2: (a), (b) presents CR and Non-CR patches (with 11 $\times$ 11 pixel resolution, 100 patches per class), from the DECam images respectively. (c) Each atom from the dictionary (121 $\times$ 256) is represented as an 11 $\times$ 11 image patch.

TABLE I: Quantitative findings on CR detection performance with the DECam test data. The true-positive rate (TPR) is evaluated at a fixed false-positive rates (FPR) of 0.01% and Precision is evaluated at a fixed TPR of 95%.

| **Algorithm** | **loss in performance** | |
|---|---|---|
| | **TPR at 0.1% FPR** | **Precision at 95% TPR** |
| **DL method** | 83.15 | 92.39 |
| LACosmic | 96.22 | - |
| deepCR (2-4) | 96.18 | 94.49 |
| DL + deepCR (2-4) | 98.16 | 97.26 |
| deepCR (2-8) | 98.74 | 98.26 |
| DL + deepCR (2-8) | 98.21 | 98.17 |
| deepCR (2-16) | 98.79 | 98.70 |
| DL + deepCR (2-16) | 99.02 | 99.04 |
| deepCR (2-32) | 99.57 | 99.01 |
| DL + deepCR (2-32) | 99.61 | 98.94 |

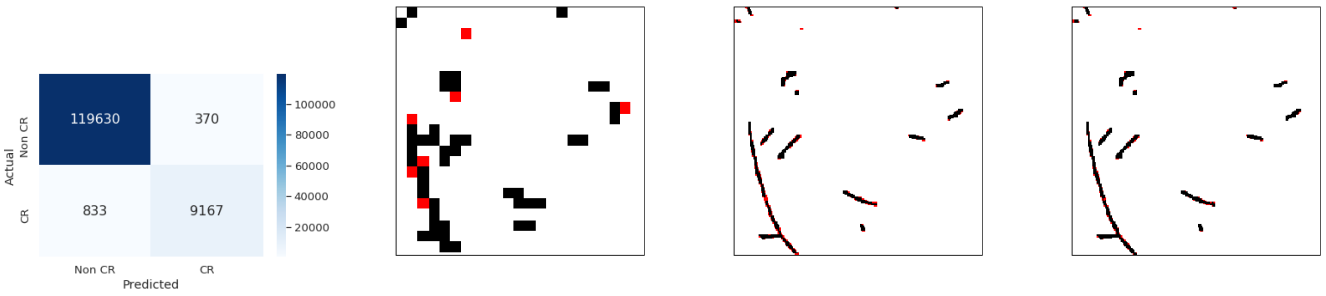
### A. Dictionary Learning Performance

First, we created a set of non-overlapping CR and Non-CR patches with 11 $\times$ 11 resolution using training images from the DECam dataset. Fig 2 illustrates the CR and Non-CR patches extracted from the DECam data. After the patches are extracted, they are converted to 121-dimensional vectors. A dictionary is then learned using 50K Non-CR patches similar to the method described in Section IV-A. The dictionary is trained with 256 atoms using the Approximate K-SVD algorithm. We experimented with multiple patch sizes and found 11 $\times$ 11 best distinguishes the CR and Non-CR patches. The number of atoms is chosen to be 256 so that it is more



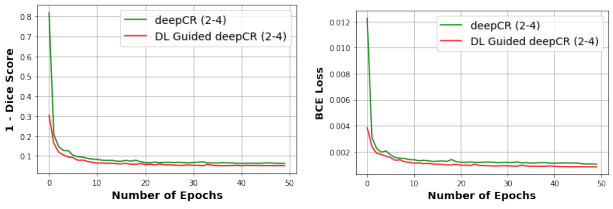

(a) DL - ROC (b) DL - PRC (c) deepCR - ROC (d) deepCR - PRC

Fig. 3: (a) and (b) presents the ROC and PRC plots obtained with the proposed DL algorithm on DECam test data. Similar plots for the deepCR (2-4) model with and without DL guidance are in (c) and (d).




(a) DL - Confusion Matrix (b) DL - Mask (c) deepCR - Mask (d) DL + deepCR - Mask

Fig. 4: CR detection performance on DECam test data with DL algorithm is presented with confusion matrix and output CR mask on the image from Figure 1 in (a) and (b), respectively. (c) and (d) are the predicted CR masks with deepCR (2-4) models with and without coarse CR map augmentation. Missing or incorrect CR pixels are marked in red.




(a) Epoch vs Performance (b) Epoch vs Loss

Fig. 5: Gains with DL-guided deep CNN models. (a) Performance using 1 - Dice Score, and (b) Binary Cross Entropy (BCE) loss over the training epochs.

than twice the dimension of the input data vector, which is 121. The dictionary atoms learned on Non-CR patches are illustrated from Figure 2. Once the dictionary is learned, both CR and Non-CR patches are represented as 256-dimensional vectors with 12 non-zero coefficients using this dictionary. We then consider another dataset from the DECam training data with 332.8K patches (25.6K are CR and 307.2K are Non-CR) for training the RF classifier to discriminate between CR and Non-CR patches. Test data consisting of 130K patches (with 10K CR and 120K Non-CR) are used for evaluating the proposed DL-based algorithm. The performance is presented with Receiver Operating Characteristic (ROC) and Precision-Recall Curve (PRC) plots in Fig. 3, Fig. 4 and Table I both quantitatively and qualitatively. We demonstrate that the DL-based algorithm can detect the CR hits at the patch level and provide approximately 83 % True Positive Rate (TPR) at 0.1 % False Positive Rate (FPR).

### *B. Deep Learning Performance*

The original deepCR is a two-layer U-Net model composed of 32 filters in the first convolutional layer (deepCR (2-32)). To evaluate the efficacy of the augmented input on the model complexity, we experimented with the number of filters in the first convolutional layer of the network. Specifically, we chose 4, 8, 16 and 32 filters for this experiment. Following the standard U-Net architecture, the number of filters in the first layer, in turn, affects the number of filters in the subsequent layers of the model. To facilitate batch training, we converted the training images from the DECam dataset to image chinks of 256 $\times$ 256. We trained the models with 20480 chunks, of which 1% (2048) are reserved for validation. We used test images from the DECam dataset with 16 images from each band for evaluation. The experimental results on the deepCR model with and without augmenting the coarse CR maps obtained from the DL algorithm are illustrated in Table I, Fig. 3 and Fig. 4. From these quantitative findings in Table I, we demonstrate that the DL-guided CNN models either outperform or match the baseline models. deepCR 2-4 provide 96 % TPR, and the same model with augmented coarse CR maps provides 98 % TPR at a fixed 0.1 % FPR. Thus, the

coarse CR augmentation from DL is helping in cases with shallow models more than deep ones. Further, the DL guidance is also helping in faster and better convergence rate shown in Figure 5 with loss and performance over training epochs.

## VI. Conclusions

In conclusion, we demonstrate the efficacy of the proposed DL-based algorithm to detect the CR hits at patch level on the DECam imager. Even though the performance of the DL method is relatively lower than the CNN-based models, this is obtained with only $256 \times 121$ parameters. We came up with this approach motivated by the unique spatial signatures of the CR hits and the natural properties of the image. In addition, the output from the DL algorithm, a coarse CR segmentation map, is augmented with the input image before feeding to the CNN models. The efficacy of the augmented input on the model complexity is evaluated by changing the number of filters in the first convolutional layer of the network. We demonstrate that the DL-guided CNN models either outperform or match the baselines in CR detection. The guidance from DL is specifically helpful for the shallow models than the deep models. Finally, the DL guidance also leads to faster and better convergence rates than the corresponding baseline models.

## VII. Acknowledgment

This work was supported by TCS and DST-ICPS (T-641). This project used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaboration. Funding for the DES Projects has been provided by the DOE and NSF (USA), MISE (Spain), STFC (UK), HEFCE (UK), NCSA (UIUC), KICP (U. Chicago), CCAPP (Ohio State), MIFPA (Texas A&M), CNPQ, FAPERJ, FINEP (Brazil), MINECO (Spain), DFG (Germany) and the Collaborating Institutions in the Dark Energy Survey, which are Argonne Lab, UC Santa Cruz, University of Cambridge, CIEMAT-Madrid, University of Chicago, University College London, DES-Brazil Consortium, University of Edinburgh, ETH Zürich, Fermilab, University of Illinois, ICE (IEEC-CSIC), IFAE Barcelona, Lawrence Berkeley Lab, LMU München and the associated Excellence Cluster Universe, University of Michigan, NOIRLab, University of Nottingham, Ohio State University, OzDES Membership Consortium, University of Pennsylvania, University of Portsmouth, SLAC National Lab, Stanford University, University of Sussex, and Texas A&M University.

## References

[1] Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al., "Lsst: from science drivers to reference design and anticipated data products," *The Astrophysical Journal*, vol. 873, no. 2, pp. 111, 2019.

[2] A Popowicz, AR Kurek, T Blachowicz, V Orlov, and B Smolka, "On the efficiency of techniques for the reduction of impulsive noise in astronomical images," *Monthly Notices of the Royal Astronomical Society*, vol. 463, no. 2, pp. 2172–2189, 2016.

[3] S. Desai, J. J. Mohr, E. Bertin, M. Kümmel, and M. Wetzstein, "Detection and removal of artifacts in astronomical images," *Astronomy and Computing*, vol. 16, pp. 67–78, July 2016.

[4] Michal Aharon, Michael Elad, and Alfred Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[5] Simon Beckouche, Jean-Luc Starck, and Jalal Fadili, "Astronomical image denoising using dictionary learning," *Astronomy & Astrophysics*, vol. 556, pp. A132, 2013.

[6] Tiep Huu Vu and Vishal Monga, "Fast low-rank shared dictionary learning for image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5160–5175, 2017.

[7] Mettu Srinivas, Yen-Yu Lin, and Hong-Yuan Mark Liao, "Deep dictionary learning for fine-grained image classification," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 835–839.

[8] Keming Zhang and Joshua S Bloom, "deepcr: Cosmic ray rejection with deep learning," *The Astrophysical Journal*, vol. 889, no. 1, pp. 24, 2020.

[9] James E Rhoads, "Cosmic-ray rejection by linear filtering of single images," *Publications of the Astronomical Society of the Pacific*, vol. 112, no. 771, pp. 703, 2000.

[10] Wojtek Pych, "A fast algorithm for cosmic-ray removal from single images," *Publications of the Astronomical Society of the Pacific*, vol. 116, no. 816, pp. 148, 2003.

[11] L Shamir, "A fuzzy logic-based algorithm for cosmic-ray hit rejection from single images," *Astronomische Nachrichten: Astronomical Notes*, vol. 326, no. 6, pp. 428–431, 2005.

[12] Pieter G Van Dokkum, "Cosmic-ray rejection by laplacian edge detection," *Publications of the Astronomical Society of the Pacific*, vol. 113, no. 789, pp. 1420, 2001.

[13] Curtis McCully and Malte Tewes, "Astro-scrappy: Speedy cosmic ray annihilation package in python," *Astrophysics Source Code Library*, pp. ascl–1907, 2019.

[14] FD Murtagh and HM Adorf, "Detecting cosmic-ray hits on hst wf/pc images," in *European Southern Observatory Conference and Workshop Proceedings*, 1991, vol. 38, p. 51.

[15] Steven Salzberg, Rupali Chandar, Holland Ford, Sreerama K Murthy, and Richard White, "Decision trees for automated identification of cosmic-ray hits in hubble space telescope images," *Publications of the Astronomical Society of the Pacific*, vol. 107, no. 709, pp. 279, 1995.

[16] Maxime Paillassa, Emmanuel Bertin, and Hervé Bouy, "Maximask and maxitrack: Two new tools for identifying contaminants in astronomical images using convolutional neural networks," *Astronomy & Astrophysics*, vol. 634, pp. A48, 2020.

[17] Chengyuan Xu, Curtis McCully, Boning Dong, D Andrew Howell, and Pradeep Sen, "Cosmic-conn: A cosmic ray detection deep-learning framework, dataset, and toolkit," *arXiv preprint arXiv:2106.14922*, 2021.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[19] B. Flaugher et al., "The Dark Energy Camera," *Astron. J.*, vol. 150, pp. 150, 2015.

[20] S. Desai et al., "The Blanco Cosmology Survey: Data Acquisition, Processing, Calibration, Quality Diagnostics and Data Release," *Astrophys. J.*, vol. 757, pp. 83, 2012.

[21] S. Desai, J. J. Mohr, R. Henderson, M. Kümmel, K. Paech, and M. Wetzstein, "CosmoDM and its application to Pan-STARRS data," *Journal of Instrumentation*, vol. 10, no. 6, pp. C06014, June 2015.

[22] Emmanuel J Candès and Michael B Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[23] Ron Rubinstein, Michael Zibulevsky, and Michael Elad, "Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit," Tech. Rep., Computer Science Department, Technion, 2008.

[24] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar conference on signals, systems and computers*. IEEE, 1993, pp. 40–44.

[25] Frederick Livingston, "Implementation of breiman's random forest machine learning algorithm," *ECE591Q Machine Learning Journal Paper*, pp. 1–13, 2005.